前沿物理大科学装置数据策略的一些思考

陈 刚

中国科学院高能物理研究所 北京 100049

摘要 前沿物理大科学装置是物理基础研究和满足国家战略需求的国之重器。特别是近年来，我国加大对前沿物理大科学装置建设与运行的投入，为基础及应用研究提供了重要的平台。大科学装置产生的数据对科学计算、数据分析、数据管理提出巨大的挑战。除了需要关注建设高水平的数据处理和科学计算平台所需要的软硬件技术以外，应同时关注极大发挥数据作用所需的策略问题。文章分析了前沿物理大科学装置的数据特点，就数据的共享、数据长期保存及再利用、数据人才策略3个方面进行简要的讨论，希望对我国大科学装置的数据应用提供参考。

关键词 大科学装置，大数据，数据管理，数据共享，数据保存，前沿物理

DOI 10.16418/j.issn.1000-3045.2018.08.015

前沿物理大科学装置占我国大科学装置比例最大，对国家科学研究和国家战略需求尤为重要。这类大科学装置产生的数据规模最大、数据结构最复杂。如何高效地发挥这些数据的最大效益是我们追求的目标。如何将数据在科学研究中的应用不是本文的重点。本文试图对数据的管理与应用中涉及的部分政策保障等作简要讨论，希望能为数据生产者、使用者、大科学装置投资者在制定政策和策略时提供参考。

1 前沿物理大科学装置简介

前沿物理大科学装置是当今国内外基础物理及应用科学研究最重要的手段和条件，包括大型粒子物理实验装置、中微子实验、重离子加速器、托卡马克实验、FAST 和 LAMOST 等天文望远镜、地面及太空宇宙线与天体物理观测装置、同步辐射实验平台、散裂中子源实验、稳态强磁场实验装置等[1]。粒子物理实验、聚变实验、天文望远镜及宇宙线观测装置基本属于专用装置，用于从微观到宇观尺度研究物质基本结构及宇宙演化等前沿科学问题。同步辐射、散裂中子源及稳态强磁场实验装置属于国家公共实验平台，向科研及产业界开放用于生命科学、材料科学、化学、物理学等领域的微观研究及高新技术开发。

国际高能物理实验以欧洲核子中心的大型强子对撞机 LHC 实验[2]为代表，每年产生的数据达数十 PB。北京

修改稿收到日期：2018年8月14日

正负电子对撞机是中国最重要的高能物理实验装置，近年来产生的数据达到 10 PB 以上。宇宙线与天体物理观测平台大致分地面宇宙线观测平台及空间科学卫星两大类。羊八井宇宙线观测站和正在稻城建设的大型高海拔宇宙线观测站 LHAASO 是国际上最重要的地面观测站，每年采集的宇宙线数据将达到 PB 量级。

我国的同步辐射光源包括运行的北京同步辐射光源、上海同步辐射光源、合肥同步辐射光源，以及（即将开工的）北京高能光源和（在建的）上海硬 X 射线自由电子激光装置；此外，中国散裂中子源已经建成投入运行。这些公共实验平台每年将吸引来自各学科领域的数千名科学家开展实验，产生的数据也达到 PB 量级。所有大科学装置产生的海量数据都是科学研究的第一手资料，是产生科学成果的源泉。

2 数据共享与利用

前沿物理大科学装置的特点是装置规模大，建设和运行周期长，其科学技术目标为瞄准国际科学技术前沿，为国家经济建设和社会发展作出战略性、基础性和前瞻性贡献。前沿物理大科学装置产生的数据是产生科学成果的金矿。装置性质的不同，数据共享与应用的模式也不同。

（1）粒子物理实验（包括对撞机实验、中微子实验、宇宙线实验等）。当前我国科学家参与的粒子物理实验包括以国外为基地的实验，如欧洲核子中心的 LHC 实验，以及以中国为基地且中国主导的实验，如北京正负电子对撞机 BESIII 实验，大亚湾中微子以及高海拔宇宙线观测站 LHAASO 实验等。所有这些粒子物理实验均采用国际合作的模式，合作各方共同分担实验的建设、运行及管理的任务。因此，粒子物理实验数据基本采用合作成员单位内自由共享、共同利用的模式，合作组内产生的科学成果以集体名义发表并共同拥有成果。尽管如此，大型粒子物理实验的合作成员之间存在竞争，合作各方都尽最大的努力争取首先获得研究成果，提升自己在合作组内及国际上的显现度。因此，除了投入最优秀的科学家以外，需要在数据传输、计算条件方面创造良好的条件，以便以最快的速度产生科学成果。中国在 LHC 实验的建设与升级方面作出了重要的贡献，但是在数据传输共享及科学计算方面的投入不足，这对中国科学家开展 LHC 物理研究造成不利影响。由于 LHC 开始向高亮度升级，数据产生率将有数十倍的增长，这对数据的传输和处理提出巨大的挑战。建议国家在网络及分布式计算等方面给予 LHC 实验中国组强有力的支持，促进中国科学家利用 LHC 国际合作实验数据产生一流物理成果。同时，我们在以中国为主导的粒子物理实验中具有管理主动权。在公平合作的前提下，我们应采取适当的策略和技术手段，在数据共享和利用方面取得主动权。

（2）天文观测（特别是大型通用型望远镜）。在国际上，该研究领域数据的共享大部分采用延时公开。天文观测者在望远镜上取得的观测数据经过一段保护期后将公开发布。在保护期内，观测者可以独享数据并尽快进行数据分析以获得科学成果。保护期后，数据将存放在数据库服务器上供世界各国的天文学家访问和使用。一般这种延时为 1—2 年。天文观测数据的这种共享方式值得其他领域学习。一方面，数据的公开可以让更多的科学家充分利用数据获得更多的研究成果；另一方面，把数据交给同行更有利于检验自己的成果。目前，空间科学卫星及宇宙线观测实验也借鉴这种模式，以一定的方式将卫星观测数据和宇宙线观测数据分批公开，提供给同行用于科学研究。

（3）同步辐射装置及散裂中子源装置。此类装置是国家投资建设的公共实验平台。学术领域的科学家可申请在平台上开展实验，实验产生的数据将用于科学研究。国外的同步装置对实验数据有相应的政策[3]。欧洲同步辐射光源（ESRF）规定，ESRF 将保存所有实验的原始数据和元数据。数据有为期 3 年的保护期，必要时可以延长。在保护期内，实验者有完全的使用权。保护期

过后，ESRF 根据相应的许可条件下将数据向 ESRF 的注册用户公开。用户使用数据产生研究成果在发表时须标明引用。国内的同步辐射装置及散裂中子源装置为大学及研究机构的科学家开放免费使用。目前，国内这些装置还没有统一的数据政策，这不利于发挥实验数据的最大利益。由于这些装置都是国家投资建设和运行，国家对装置产生的数据应该拥有共同所有权。因此，建议国家建立与国际上类似的数据政策，既保护实验者对数据的优先使用权，也通过数据共享充分发挥数据的作用。公共实验平台的数据共享可以采取两种模式：① 建立数据保护期，期限 2—3 年，确保实验者对数据的优先使用权。② 对急需使用实验数据的外部用户，可以与实验者签订合作协议，建立数据共享机制，让这些用户在保护期也能及时利用数据开展科学研究。

3 数据保存及再利用

前沿物理大科学装置的建造、维护和数据采集消耗了大量的人力、物力，因此实验数据是极其宝贵的。科学家对数据的利用不会随着数据采集的结束而立即停止，很多实验在数据采集结束后的若干年内，仍然在进行数据分析研究并有相关的论文发表。不同大科学装置的数据具有唯一性，随着理论研究的进步和分析手段的提高，旧的实验数据中可能会有新的科学发现。另外，对不同实验的新、旧数据的联合分析和交叉验证，能够提高科学发现的精度和可信度。大科学装置的数据的另一个重要用途是提供给大专院校和中小学校用于教学和科普。由此可见，前沿物理大科学装置的数据的长期保存具有极其重要的意义。

数据的保存不仅仅是实验采集的数据，还应该包括知识库。所谓知识库包含描述实验条件的参数、分级数据所用的软件、文档以及其他数据分析所需的资料。所以这些信息是保证正确进行数据再利用和分析的必要条件。后续数据分析的类型不同，有些数据分析需要使用实验的原始数据，有些只需使用经过处理的高级数据，这对数据的长期保存提出不同要求。以高能物理为例，国际高能物理领域成立了数据长期保存合作组 DPHEP（中国科学院高能物理研究所是发起单位之一），并编写了数据长期保存技术白皮书[4]。该白皮书对数据和知识库的保存、相关技术及策略进行了详尽的描述。我国前沿物理大科学装置的数据策略缺乏系统的数据长期保存及再利用的规划及策略，因此该白皮书对国家制定相关政策具有很好的参考意义。此外，我国的经费资助基本是针对项目的，当大装置运行结束后，很难得到对数据保存给予支持的后续经费。因此，应建立相应的资助机制，以确保大科学装置运行结束后数据的长期保存和高效再利用。

4 人才策略

前沿物理大科学装置是目前中国规模最大的一批科学装置，产生的数据规模也是空前的。管理和分析这些数据需要最先进的算法和软件。这对人才队伍提出了巨大的挑战。数据分析的算法和软件一般需要相关物理专业的人才进行开发和实现。而大部分物理专业人员在计算机技术方面训练不够，特别是年轻硕博士毕业生和博士后在工作中将面临数据分析工具、软件及编程语言等问题。因此，一个大科学装置项目应该为这些物理学专业人员提供在职的计算机技术培训。欧洲核子中心每年举办高水平的计算技术暑期学校[5]，挑选世界各国的优秀青年学生或青年科学家参加，提供科学计算技术培训课程和实习。国内应针对不同的大科学装置或者不同的科学计算方法建立高水平的培训课程，鼓励科学家参加计算技术培训。这将极大地推动科研人员软件及数据分析的水平，促进科学产出。

前沿物理大科学装置的规模和复杂度都是空前的。数据分析的过程复杂、任务量巨大，仅靠物理学专业的科学家是不够的。另外，在光源、散裂中子源等公共实验平台上开展实验的科学家来自不同的专业领域，对实验平台的结构及数据结构了解不一定深入。这种情况下，计算机专业的科研技术人员与物理学及其他相关领域科学家的合作将变得非常重要。以中国科学院高能物理研究所为例，

其组建了一支计算机专家队伍与物理学家配合和沟通，对数据分析软件进行优化；同时，物理学家依据物理分析计算的需求和特性与计算机专家深入沟通，对计算机硬件平台、数据管理系统、中间件系统进行优化。在物理学家开发数据分析软件时，计算机专家帮助他们优化软件，提高数据访问的效率和软件运行的效率。计算机专家同时根据物理学家访问数据的特点和对 CPU 的利用特点设计和建造数据存储系统以及计算集群的体系结构，使数据处理达到最高的效率。双向沟通确保数据分析计算系统能以最高效率满足科学计算的要求。

高水平的软件是大科学装置成功的关键。软件开发人员的工作得到认可并在职业晋升、待遇等各方面得以保障是吸引高水平软件人才稳定从事大科学装置数据与计算软件开发及运行维护的必要条件。以高能物理为例，欧洲核子中心拥有一支高水平的计算机及物理软件开发的队伍。这支队伍几十年来专心进行大型通用物理软件及数据分析软件的研究发展，其产生的“www 技术”成为全世界最重要的网络技术，极大地推动互联网的发展。另外，欧洲核子中心开发的物理模拟软件“GEANT4”成为全球粒子物理、核物理、核医学以及射线技术计算的基础，“ROOT”成为数据分析的核心技术。这说明确保科学家全心全意投入软件研究开发的重要性。一方面，为提升软件开发者的显现度，应鼓励他们将软件开发的技术和成果写成文章发表。另一方面，应该鼓励或要求领域科学家在发表文章和论文时恰如其分的引用其使用的软件。这对正确认可软件开发者的贡献尤其重要。

参考文献

1 中国科学院条件保障与财务局. 中国科学院重大科技基础设施. [2018-08-14]. http://lssf.cas.cn/dzzRegisterController.do?outerelss&flag=99.

2 European Organization for Nuclear Research (CERN). Large Hadron Collider. [2018-08-14]. https://home.cern/topics/large-hadron-collider.

3 European Synchrotron Radiation Facility. ESRF Data Policy. [2018-08-14]. http://www.esrf.eu/datapolicy.

4 The DPHEP Study Group. Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. [2018-08-14]. https://arxiv.org/pdf/1205.4667v1.pdf.

5 European Organization for Nuclear Research (CERN). CERN School of Computing. [2018-08-14]. http://csc.web.cern.ch.

Perspective on Data Strategy for Large Facilities of Physics Frontiers

CHEN Gang

(Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China)

Abstract Large research facilities play important roles in the fundamental and applied scientific researches in China. In recent years, the investment has been increasing in large facilities of physics frontiers which provides essential infrastructure of physics researches. These large facilities lead to big challenges to scientific research, data analysis and management. In addition to the computer technologies to build the high performance computing platform, the strategy to use and manage the data may be more important to guarantee the large facilities have the most scientific merits. Trying to facilitate the scientific productivity, this article will discuss the strategies of data sharing, data preservation and reuse, staffing and careers based on the characteristics of data generating from different facilities.

Keywords large facilities, big data, data management, data sharing, data preservation, physics frontier



陈 刚 中国科学院高能物理研究所副所长，研究员，博士生导师。1982年毕业于南京大学物理系。1994年毕业于中国科学院高能物理研究所并获得博士学位。20世纪90年代，参与欧洲核子中心（CERN）的L3大型高能物理实验和阿尔法磁谱仪AMS项目，以及北京正负电子对撞机上的BES实验。2003年起，主持建设用于高能物理实验的高性能计算平台；期间与欧洲核子中心合作，参与建设高能物理网格，为LHC等大型高能物理实验提供计算支撑。2005年任国际高能物理计算协调委员会委员。2006年起，参与FP6/FP7框架关于网格及云计算合作项目建设，担任中方协调人以及项目管理委员会主席。2017年牵头承担科技部重点研发计划项目“面向高能物理领域科学发现的高性能应用软件系统研制”。E-mail: Gang.Chen@ihep.ac.cn

CHEN Gang Professor, Deputy Director of Institute of High Energy Physics (IHEP), Chinese Academy of Sciences (CAS). He got his BSc degree from Nanjing University in 1982 and his Ph.D. from IHEP in 1994. In 1990s, he worked as an experimental physicist on L3 experiment at CERN, Alpha Magnetic Spectrometer (AMS), and BES experiment on Beijing Electron Positron Collider (BEPC). Since 2003, he has been in charge of the provision of a high performance computing infrastructure for high energy physics projects. In 2005, he became the member of IHEPCCC. Since 2006, he has been the leader of EU FP6/FP7 projects on grid and cloud computing. In 2017, he became the Principal Investigator of the Project of High Performance Application Software System for High Energy Physics (HEPHPC), an HPC project of National Key R&D Plan of Ministry of Science and Technology. E-mail: Gang.Chen@ihep.ac.cn

■责任编辑：岳凌生